The combination of the Stuyfzand-classification and Fuzzy C-Means clustering in a Hydrogeochemical System Analysis of the IJssel valley, The Netherlands (a case study)

S.J.A. van Baalen
Institute for Inland Water management and Waste Water Treatment (RIZA), P.O. box 17, 8200 AA Lelystad, The Netherlands

J.C. de Ruiter
Drinking Water Company of Overijssel (WMO), P.O. box 10005, 8000 GA Zwolle, The Netherlands

In the eastern part of The Netherlands problems arise with the production of good quality drinking water because of upconing relict salt water underlying the deep fresh water aquifer.
Instead of making a laborious hydrological modelling, we choose to make a hydrogeochemical system analysis on a regional scale. Based on the results of this analysis, a decision on the necessity of a more detailed hydrological modelling could be made.
The objective of the hydrogeochemical system analysis is to find and characterize, using a qualitative analysis, different types of groundwater and their origin. Classification of water types is the main theme. This classification was made by means of the Fuzzy C-Means clustering and Non-Linear Mapping, in combination with the classification of water types according to Stuyfzand [13,14].
With the results of the classification, the regional pattern of groundwater quality could be described very well, and the cause of the salinization could be determined.
The use of ***FCM in combination with the Stuyfzand-classification*** *is very much recommended for the analysis of hydrochemical data, as these methods make it possible to get a good and interpretable picture of a complex matter such as chemical analyses scattered over many dimensions.*

Introduction

A hydrological system analysis, as described by Engelen & Jones [6], is an excellent way for getting insight into a regional hydrological system. The main aim of a hydrological system analysis is the mapping of nested groundwater flow systems of different orders, each connecting a recharge area with one or more discharge areas. When executing a hydrological system analysis, different kinds of interrelated data are used, such as geological (formations, sedimentology), (geo)hydrological (properties, flow patterns), topographical (surface levels, landuse), geochemical (soil) and hydrochemical (groundwater) data.

In the present study, the emphasis is on geohydrochemical interpretation. The problem with hydrochemical data however, is how to describe the properties of one sample point in a simple but meaningful way. Matthes [9] distinguishes about 10 different ways that can be used for solving this problem. However, as pointed out by Stuyfzand [15], each method is intended for a specific purpose and has its own merits, but all of them lack a more "holistic view on water quality". Therefore Stuyfzand [13, 14, 15] developed a procedure to map and diagnose the major factors accounting for regional variations in the hydrochemistry. In this procedure, a classification system to determine water types plays an important role.

An alternative way of classifying water samples is to make use of statistical methods. If any *a priori* knowledge of the processes responsable for the inhomogenity of the dataset is known, conventional multivariate statistics like discriminant function analysis can be used [4, 20, 21]. When there is no such *a priori* knowledge, conventional cluster techniques are commonly used to reorganize the data set into homogeneous groups, which can only be interpreted afterwards [8, 20, 21].

Both discriminant function analysis and conventional cluster techniques have the disadvantage of being rigid, i.e. they assign a sample unambiguously to a group or cluster [8]. However, the spatial variation in hydrochemical data is usually gradual instead of abrupt. Therefore, some overlap between groups is needed to describe the hydrochemical pattern in a sufficient way. A method using fuzzy or continuous classes is thus more appropriate, as it allows overlap between classes [1, 5, 20].

In this paper, we make a comparison between the classification system proposed by Stuyfzand [13, 14, 15] and the Fuzzy C-Means clustering in combination with Non-Linear Mapping used by Vriend et al. [20] and Frapporti et al. [8]. It is not our intention to qualify one of these methods as best, but we want to show how these methods can supplement each other to get a better result than with either single method.
After a brief introduction into the theoretical background of the two methods, a case study in the eastern part of The Netherlands is used to show the applicability of this combination. In the end, some general conclusions will be drawn.

Theoretical Background

In this paper we only summarize the relevant information of both methods. A complete overview is found in the original publications and the handbooks mentioned in the list of references [1, 5, 8, 13, 14, 15, 20].

The classification of water types according to Stuyfzand [13,14]
The classification system combines four essential aspects in a logical code (see Figure 1): the chlorinity (main type), alkalinity (type), most important cation and anion (subtype) and a base exchange index (class). By calculating these four aspects of a certain water sample, the water type is determined in a hierarchical way.

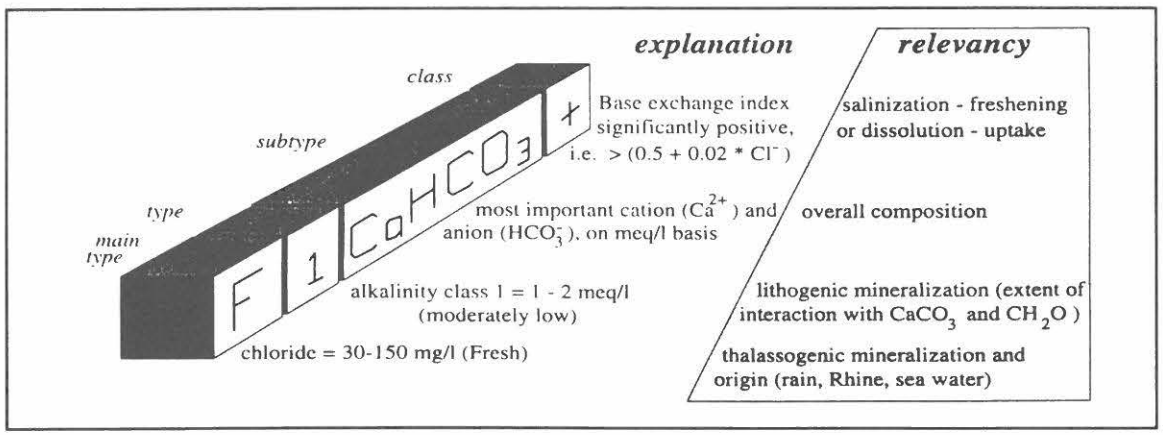

Figure 1: *The hydrochemical classification system of water types by Stuyfzand [14], with its coding in 9 positions. The example is called "a fresh, moderately low alkalinity, calciumbicarbonate water with a positive base exchange index" [15]*

The main advantage of this classification system is that diagnosis of different chemical processes is possible and that the resulting water type of a sample is always the same, because of the predetermined boundaries. Its differentiating power is however in principle always the same, because of those predetermined boundaries. Stuyfzand recognizes the stiffness of the classification system and proposes ways to associate or differentiate within the system. Association takes place by combining two or more main types or by leaving away class and/or subtype. Differentiation can be done by dividing some levels in a more subtle way or by adding new parameters. This flexibilisation however, is still based on presuppositions.

The Fuzzy C-Means clustering in combination with Non-Linear Mapping [8, 20]
This clustering is a powerful tool for separating datasets in homogeneous groups, without any *a priori* knowledge. The technique allows some vagueness in the description of the cluster model. Likeness or similarity of a sample with a cluster group is indicated by a continuous function (membership) between zero (completely different) and one (exactly the same). The memberships of a case to all clusters sum up to 1. Hence, it is possible for a sample to "belong" to more than one cluster, which allows some flexibility in the interpretation of the data set. This flexibility is an important difference with so-called "hard" statistical techniques.
An important issue in executing this clustering is the choice of the number of clusters. Several runs [19] are made, each time with a different number of clusters. A definite choice is based on various criteria. In the present study, our criteria were based on the regional distribution, the hydrochemical interpretability and the "structure" as displayed in a two-dimensional image made by Non-Linear Mapping [8].
The Non-Linear Mapping determines a 2- or 3-dimensional image of a X-dimensional data cloud, such that the interdata distances are minimally distorted [20].
Although the compositional overlap of the different cluster groups is an advantage while interpreting the results, for presentational purposes some "hardening" of the clustering is needed. This "defuzzication" takes place based on the criterium that a sample point is assigned to a certain cluster if the highest membership of that case is 1.67 times higher than the one-highest membership.

Comparison of the Stuyfzand-classification and the Fuzzy C-Means clustering.
As with all classification systems, both methods have their advantages and disadvantages.
With regard to the Stuyfzand-classification and the Fuzzy C-Means clustering, the following can be said.
The advantage of the **Stuyfzand-classification** is that the resulting water type of a sample is always the same because of the predetermined boundaries and that diagnosis of different chemical processes is possible. Its main disadvantage is its stiffness, because of these predetermined boundaries.
The advantage of **Fuzzy C-Means clustering** is that it makes homogeneous groups within the dataset without *a priori* knowledge, while the differences between the groups are relevant for the specific data set. This means that the cluster groups are sometimes coarser and sometimes finer than the water types found by the classification. The main disadvantage of the clustering method is its lack of direct interpretability and its dependency of the specific data set; cluster groups of different data sets are not comparable with each other.

By means of the following case study, we will show that both methods supplement each other, thus taking away the disadvantages of each single method while keeping the advantages of both.

Case Study: Hydrogeochemical System Analysis of the IJssel valley, The Netherlands

Introduction
The Drinking Water Company of Overijssel (WMO) is active in the eastern part of The Netherlands. Recently it took over a groundwater extraction site in the city of Deventer. This site has problems with an increasing chloride-content in the collected water. For example, in the period 1970 to 1993, the chloride-content rose on average from 86 mg/l to 117 mg/l.
In order to find a new, optimal structure of the well field, the WMO wanted to know more about:
- the position of the well field in the regional hydrological system
- the processes that are responsible for the increasing salinization of the well field

We decided to do a regional hydrogeochemical system analysis, with the emphasis on geology, (geo)hydrology and hydrochemistry. Geological and hydrological data were found in literature [3, 7, 10, 11, 12, 16, 17, 18, 22]. It should be mentioned, that in this study no hydrological modelling has

been done. However, one of the results of the study is a list of boundary conditions that can be used in an eventual hydrological modelling of the area.
The hydrochemical data used in this study were mostly existing data, with which we wanted to do a qualitative analysis, with classification of water types as the main theme.

Study Area
The study area is located around the city of Deventer, measuring some 32 km from west to east and 40 km from north to south. It comprises the valley of the river IJssel in between the city of Zwolle in the north and the city of Zutphen in the south. The west- and east-borders are formed by two ice-pushed ridges, respectively the Veluwe and the Sallandse Heuvelrug (see Figure 2). These ridges form the borders of the regional groundwater system of the IJssel valley.

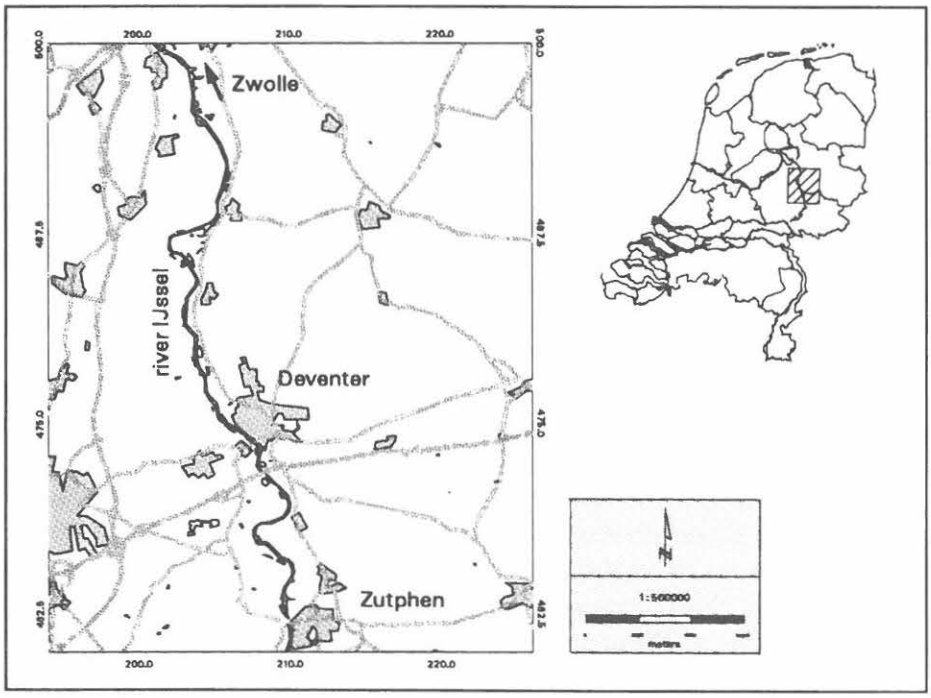

Figure 2: *Location of the study area*

Geology
In the study area, the formations containing the drinking water consist of late-Tertiary and early-Quaternary sediments (Figure 3). The older, deepest formations contain brakish-salt water; the younger formations contain fresh water.
The IJssel valley was formed during the Saale-ice age, in which the ice in The Netherlands reached its outmost extension. An ice-tongue made its way down south through the study area, pushing aside the Tertiary and Quaternary sediments. It so formed an up to 120 m deep basin, flanked by ice-pushed ridges of some 100 m. When the ice regressed, the basin filled with water. In this glacial lake, a thick packet of fluvioglacial material (basin-clay) sedimented. These sediments were during the late-Quaternary covered by fluvial and eolian deposits.

(Geo)hydrology
The two ice-pushed-ridges serve as infiltration areas while the basin-clay forms the principal division between shallow and deep aquifers in the IJssel valley. There is almost no exchange between the shallow groundwater, which is influenced by man and river, and the very old (up to 20.000 years), deep groundwater. Under the city of Deventer the division between fresh and salt water (150 mg/l Cl^-) is found at approximately 130 m below sea-level.

In this study, the main focus was on the deep groundwater, as there is relatively little known about it and because its importance as a source of drinking water.

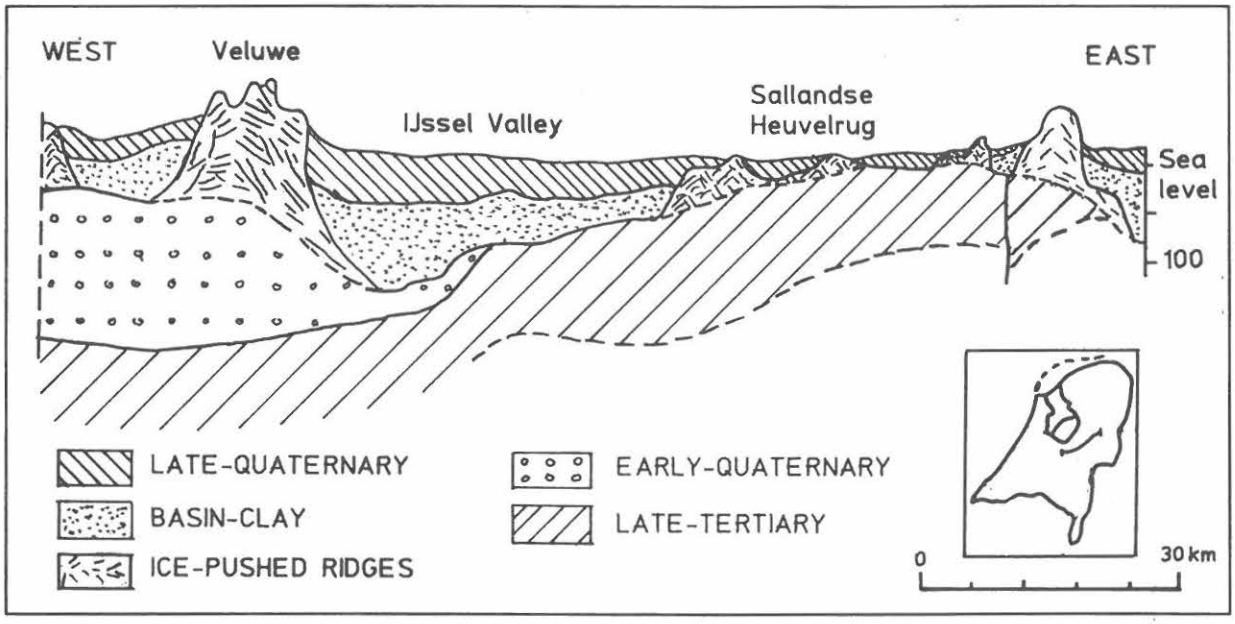

Figure 3: *Geological cross section from west to east across the IJssel valley [after 11]*

Methods of investigation

As mentioned in the introduction, we mainly used existing data. So as in the previous section is described, information about the geology and (geo)hydrology was obtained from literature. The necessary hydrochemical data were collected from the On Line Groundwater Archive (OLGA), operated by TNO-GG, and from internal WMO-archives.

The following criteria have been used by the selection of the data:

- data used should not have been collected before 1970
- analyses should contain at least Na^+, Cl^-, Ca^{2+}, HCO_3^-, Mg^{2+}, SO_4^{2-}, Fe^{2+}, NO_3^{2-}, pH and EC
- only data from monitoring wells with at least one filter below 50 m -N.A.P. (= the minimal depth of the lower side of the basin-clay formation) were used

This selection resulted in a data set of 189 analyses. 17 New analyses were added to this, resulting in a total dataset of 206 analyses located at 33 sites.

The water type was calculated using the program *ChemProc v.4.0* [2]. The clustering was executed using the program *FNX* [19].

Results and discussion

In this study, all elements apart from pH, which is intrinsically logarithmic, were lognormally distributed and therefore logarithmically transformed. Three samples were left out of the analysis, because there were indications that they were unreliable. The clustering may in principle be executed with an unlimited number of parameters, but the result is not always better when using a lot of parameters because of possible correlation between two or more parameters. The parameters ultimately chosen for the clustering were Na^+, Cl^-, Ca^{2+}, HCO_3^-, Mg^{2+}, pH and K^+. They had the highest segregational power. A seven-cluster model appeared best to describe the data set. Cluster centres of the fuzzy clusters and geometric means of the hardened clusters are given in Table 1. The clusters are arranged in order of chlorinity, i.e. cluster 1 has the lowest chlorinity and cluster 7 the highest. The Non-Linear Mapping plot in Figure 4 shows the consistency of the partitioning. The samples indicated by the number "10" are intermediate, i.e. they could not be hardened because of their low memberships.

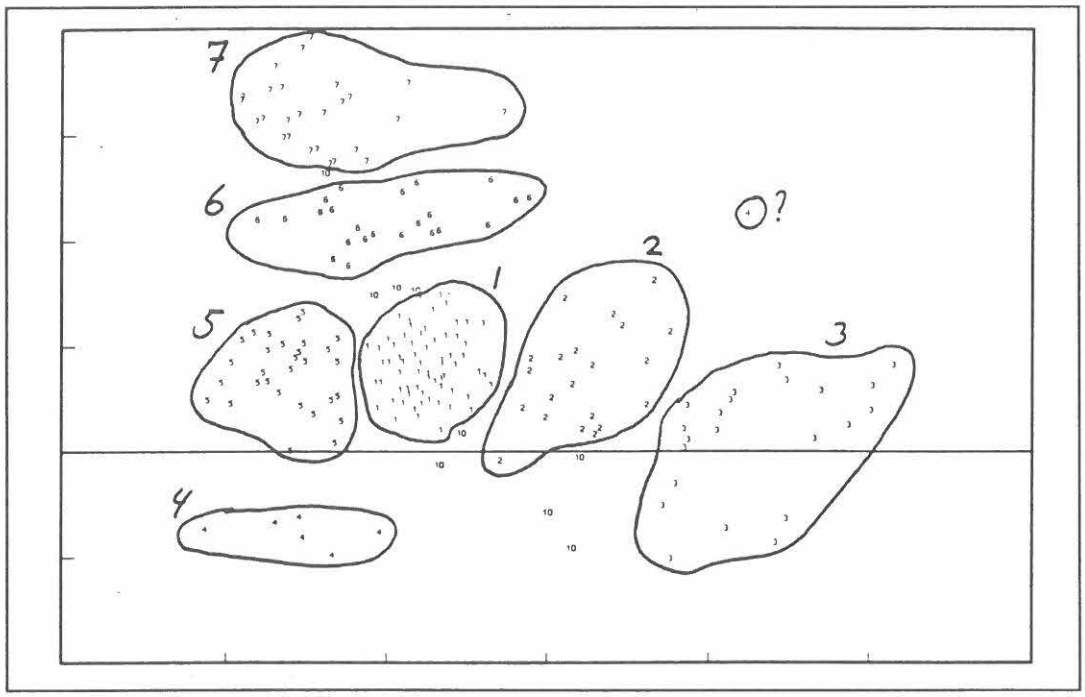

Figure 4: *Fuzzy C-Means classification represented in a Non-Linear Mapping plot. The numbers indicate the hardened cluster numbers.*

	Cluster 1 (n = 66)		Cluster 2 (n = 21)		Cluster 3 (n = 21)		Cluster 4 (n = 7)	
Variable	CC	GM	CC	GM	CC	GM	CC	GM
Na^+	21.5	23.5	14.9	17.9	11.8	30.9	20.5	52.7
Cl^-	13.5	16.3	13.6	13.7	16.3	46.9	25.4	32.1
Ca^{2+}	36.3	36.8	15.4	15.4	9.2	10.8	51.0	54.0
HCO_3^-	169.8	172.9	66.7	74.4	14.8	14.7	210.6	269.7
Mg^{2+}	4.4	4.6	2.7	2.8	1.4	3.0	7.7	8.4
pH	7.7	7.7	7.1	7.3	6.4	6.3	7.5	7.5
K^+	2.6	3.1	1.5	1.6	1.1	1.5	0.0	0.0

	Cluster 5 (n = 30)		Cluster 6 (n = 23)		Cluster 7 (n = 26)	
Variable	CC	GM	CC	GM	CC	GM
Na^+	23.7	24.4	202.4	231.4	1152.0	1290.0
Cl^-	32.6	35.3	257.0	304.0	1822.9	2068.0
Ca^{2+}	87.8	92.5	28.0	30.7	37.9	50.4
HCO_3^-	262.6	277.6	258.4	276.6	391.0	397.0
Mg^{2+}	8.9	9.3	11.6	12.3	52.0	58.0
pH	7.3	7.3	7.9	7.9	8.0	8.0
K^+	1.9	2.3	16.4	17.6	50.3	53.0

Table 1: *Cluster centers (CC) of the fuzzy clusters and geometric means (GM) of the hardened clusters for the adopted cluster model. Nine samples could not be hardened. The differences between the CC's and the CM's are a result of the effect of outliers (samples with a low membership). In calculating the CC, membership is used as weighing factor, so the effect of outliers on the CC is less than on the GM. All variables are expressed in mg/l, except pH. The total number of samples is 194 (206 minus 3 unreliable samples and 9 intermediate samples).*

A description of the clusters was made in several ways, using:

- several statistical parameters for every cluster (geometric mean, median, quartiles, minimum and maximum)
- Stiff-diagrams for the samples with the highest membership for every cluster
- the spatial distribution for every cluster
- the Stuyfzand-classification for all samples

It is beyond the scope of this article to evaluate the water type of every cluster thoroughly, only the ultimate interpretation is given. This interpretation is based on the foregoing descriptions together with information from literature. Next to the interpretation of every cluster, the spatial distribution pattern of the different clusters had to be determined. Horizontal sections at different depths and vertical cross-sections were constructed. In drawing the boundaries between the clusters, the memberships of all samples were used, together with geological data and hydrological (model)results from literature.

Table 2 gives for every cluster a short description of the interpretation of the water type, together with the dominating water type, according to the Stuyfzand-classification. This table shows that in the present study the combination of the two methods results in an association of water types according to the Stuyfzand-classification. Every cluster comprises one or two water types according to the Stuyfzand-classification. Thus, an association of water types is executed without any *a priori* knowledge.

Cluster 1 comprises the water type g2-$CaHCO_3$+: a oligohaline (rainwaterlike), moderate alkalinity, calciumbicarbonate water with a low positive base exchange index. The water in this class is rain water, infiltrated at least some 2500 years ago [3]. As a result of a complex of soil forming processes (reduction of iron and sulfate, weathering of silicates, dissolution of calciumbicarbonate) this water has changed, but is still recognizable as Old Infiltrated (rain)water (OI).

Cluster 2 comprises the water types g1-$CaHCO_3$+ and g2-$NaHCO_3$+: oligohaline, moderately low alkalinity, calcium/natriumbicarbonate water with a low positive base exchange index. The water in this class is rain water, infiltrated between the present and some 2500 years ago [3]. Because of the very low amounts of macro-ions, this water is interpreted as Young Infiltrated (rain)water (YI).

Clusters 3 till 5 comprise a mixture of different water types, ranging from very poor to very rich water, but all with high amounts of nitrate and/or sulfate. The maximum sample depth is 70 m below sea level. Agricultural and surface water influences dominate these clusters. They are therefore all interpreted as Shallow, Man-influenced water (SM).

Cluster 6 comprises the water types f3-NaCl+ and B3-NaCl+/o: fresh-brackish to brackish, moderately high alkalinity, sodiumchloride water with variable base exchange indices. Under the city of Deventer, the lowest base exchange indices point at salinization. Just outside Deventer, the high base exchange indices in this water point at freshening. These samples are interpreted as Relict Brackish water (RB).

Cluster 7 comprises the water types b3/b4-NaClo: a brackish-salt, moderately high to high alkalinity, sodiumchloride water, with a very low base exchange index. The interpretation is Relict Salt water (RS) without any freshening or salinization.

Figure 5 shows an example of a profile, crossing the IJssel valley from west to east. In this figure, the different types of water are easily recognizable and it shows that the salinization is caused by upconing salt water.

When looking at Table 2 and Figure 5, the following conclusions can be drawn:

- by using hydrochemical data together with geological and (geo)hydrological data from literature, it was possible to create a well structured picture of the different kinds of water present in the deep aquifers of the IJssel valley
- the cause and extension of the salinization in the well fields of the city of Deventer are clearly visible in the horizontal and vertical sections

FCM-cluster	Interpretation	Description	Stuyfzand-classification
1	Old Infiltrated (rain)water (OI)	Water in deep aquifers, influenced by processes like reduction and calcite dissolution	g2-$CaHCO_3$+
2	Young Infiltrated (rain)water (YI)	Slightly altered rain water in infiltration areas, lying above type 1	g1-$CaHCO_3$+ g1-$NaHCO_3$+
3,4,5	Shallow, Man-influenced water (SM)	Water above the basin-clay, strongly influenced by surface water and agricultural activities	g*-CaMiClo, g*-CaMISo (3) F2-$CaHCO_3$o (4) g2-$CaHCO_3$+ (4) F3-$CaHCO_3$o/+ (5) g3-$CaHCO_3$+/o (5)
6	Relict Brackish water (RB)	Boundary zone between deep fresh water and marine salt water	f3-NaCl+ B3-NaCl+/o
7	Relict Salt water (RS)	Marine salt water, captured in Tertiary marine sediments	b3-NaClo b4-NaClo

Table 2: *Interpretation of the different clusters, together with a short description and the dominating water type according to the Stuyfzand-classification. Note that the water types according to the Stuyfzand-classification are all unique for one Fuzzy C-Means (FCM) cluster, except g2-$CaHCO_3$+.*

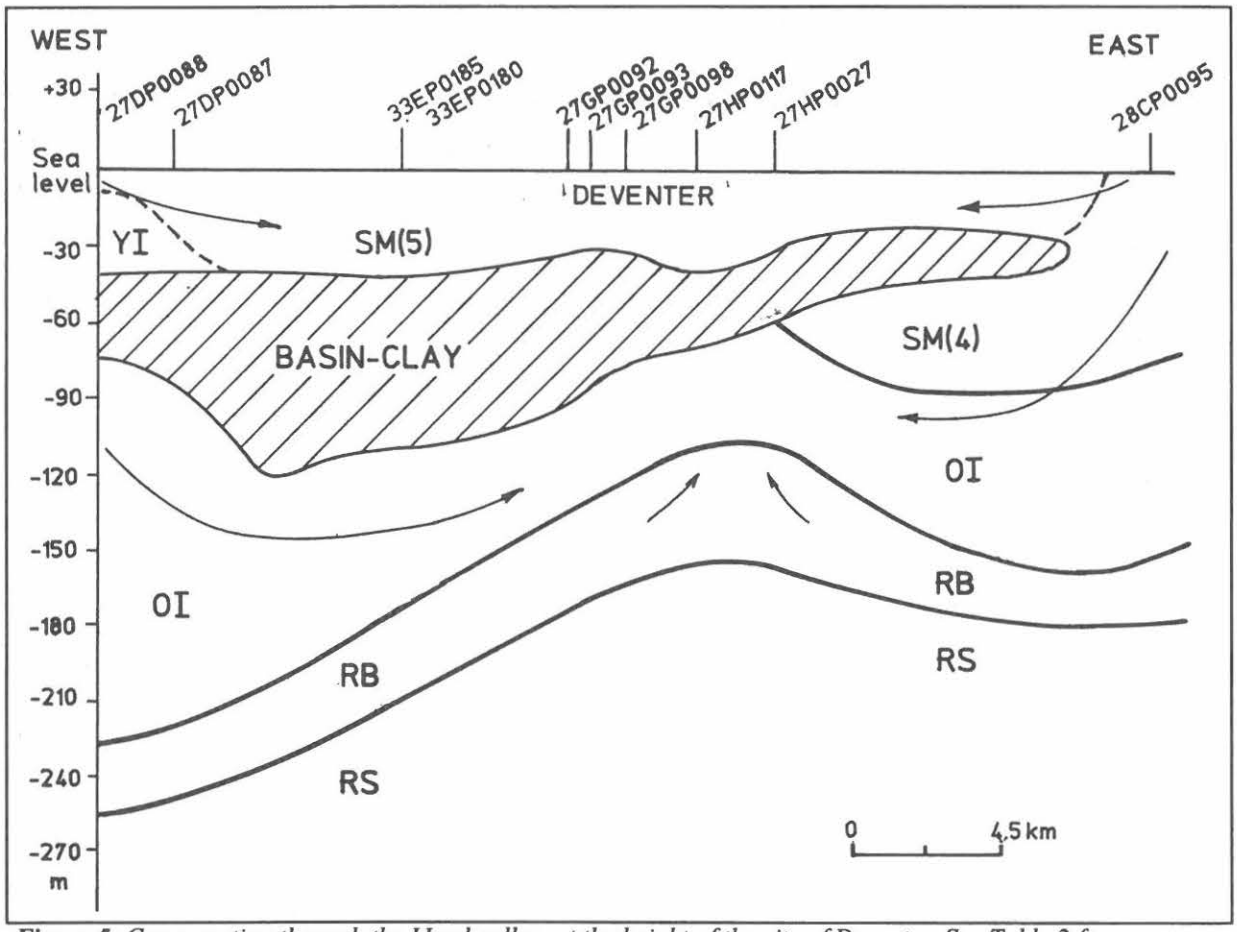

Figure 5: *Cross-section through the IJssel valley, at the height of the city of Deventer. See Table 2 for explanation of the codes.*

- It is shown in Table 2 that a combination of the clustering with the classification gives good results. The homogeneous groups within the data set, found using the Fuzzy C-Means clustering, have differences in an order of magnitude that is relevant for the data set in question. The water types according to the Stuyfzand-classification make it possible to recognize chemical processes, and to compare the clusters with other studies.

General Discussion and Conclusions

In this article, we wanted to show the applicability of the combination of the Fuzzy C-Means clustering together with the Stuyfzand-classification. The result is an easy interpretable hydrochemical classification at the appropiate level, as shown in the presented case study.
Stuyfzand [15] formulated some possible problems with applying a cluster analysis when classifying water samples. He points out that:

1. the boundaries of the clusters are not clearly defined and adjacent maps are incomparable
2. (log)normality and statistical independence have to be assumed
3. extremes have to be excluded
4. a discriminant function analysis is required to reveal the relative importance of all hydrochemical variables

There is, of course, some truth in these points, especially when already a lot is known about the specific area. But in our view the combination of the clustering and the classification method neutralizes part of these problems and is therefore very useful in a new study area with unknown processes.

ad 1. Some vagueness of the boundaries between clusters represents very well the gradual change present in nature, and is therefore greatly approved of by geographers. Adjacent cluster maps are comparable when the water types according to the Stuyfzand-classification are used.

ad 2. Multivariate (log)normality and statistical independence are not required [20] but it is good to keep these features in mind when choosing the variables for executing the clustering.

ad 3. Extremes contribute equally to all clusters and are no problem for the clustering. However, the scale of a Non-Linear Mapping plot is influenced by extremes [20] and some critical view to these extremes is still needed.

ad 4. When using water types according to the Stuyfzand-classification, no discriminant function analysis is needed to interprete the clusters.

Our conclusion is therefore that the great advantage of a statistical technique like Fuzzy C-Means clustering is its possibility to divide a data set in homogeneous groups at the appropriate level. The combination with the Stuyfzand-classification makes a good interpretation of the clusters possible.

Acknowledgements

The valuable contributions of dr. P.F.M. van Gaans are gratefully acknowledged. J. Bell, M.Sc, is thanked for her linguistic comments.

REFERENCES

[1] BEZDEK, C.J., 1981: Pattern recognition with Fuzzy Objective Function algorithms. -Plenum Press, New York

[2] BIESHEUVEL, A., 1991: ChemProc v. 4.0 -Vrije Universiteit Amsterdam, The Netherlands

[3] BROKS, 1992: Geohydrologische onderzoek winbare hoeveelheid "Salland-diep", projectno. 11.250, 4 delen

[4] DAVIS, J.C., 1986: Statistics and data analysis in geology. -John Wiley & Sons, New York, 2nd ed.

[5] DE GRUIJTER, J.J. & McBRATNEY, A.B., 1988: A modified fuzzy k-means method for predictive classification. -Classification and Related Methods of Data Analysis, H.H.Bock (ed.), Elsevier Science Publishers, The Netherlands, pp. 97-104

[6] ENGELEN, G.B. & JONES, G.P., 1986: Developments in the analysis of ground water flow systems. -IAHS Publ. 163, 365 pp.

[7] ENGELEN, G.B. et al., 1989: Grondwaterstromingsstelsels in Nederland. Ministerie van LNV, SDU-uitgeverij, Den Haag

[8] FRAPPORTI, G., VRIEND, S.P. & VAN GAANS, P.F.M. (1993): Hydrogeochemistry of the shallow dutch groundwater: interpretation of the National Groundwater Quality Monitoring Network. -Water Resources Research, Vol. 29, No. 9, pp. 2993-3004

[9] MATTHESS, G. , 1982: The properties of groundwater. -John Wiley & Sons, New York, 406 pp.

[10] RGD (National Geological Survey), 1973: Geologische opbouw van Salland en het aangrenzende randgebied, rapport no. 844 "Schalkhaar".

[11] RGD (National Geological Survey), 1975: Geologische overzichtskaarten van Nederland. Kaarten, profielen en toelichting.

[12] RGD (National geological Survey), 1990: Geologische en Hydrogeologische opbouw waterwingebied Zutphenseweg te Deventer, rapport no. BP 10804

[13] STUYFZAND, P.J., 1986: Een nieuwe hydrochemische classificatie van water typen, met Nederlandse voorbeelden van toepassing. -H_2O, Vol. 19, no. 23, pp. 562-568

[14] STUYFZAND, P.J., 1989: A new hydrochemical classification of water types. -IAHS Publ. 182, pp. 89-98

[15] STUYFZAND, P.J., 1993: Hydrochemistry and Hydrology of the Coastal Dune area of the Western Netherlands. -KIWA-report I11, Nieuwegein, The Netherlands, 366 p., also published as thesis Vrije Universiteit, Amsterdam, The Netherlands

[16] TNO-DGV, 1983: Hydrochemie van Oost-Nederland. OS 83-38, Auteur: J.H. Hoogendoorn

[17] TNO-DGV, 1986: Grondwaterstromingssystemen Salland/O-Veluwe. OS 86-36. Auteurs: G.K.Brouwer & J.H. Hoogendoorn

[18] TNO-DGV, 1990: Grondwatersysteemonderzoek Salland I en II. OS 90-48. Auteur: J.H.Hoogendoorn

[19] VAN GAANS, P.F.M. & VRIEND, S.P., 1995: FNX, a program for Fuzzy C-Means clustering and Non-Linear Mapping. -Universiteit Utrecht, The Netherlands

[20] VRIEND, S.P., VAN GAANS, P.F.M., MIDDELBURG, J. & DE NIJS, A., 1988: The application of fuzzy c-means cluster analysis and non-linear mapping to geochemical datasets: examples from Portugal. -Applied Geochemistry, Vol. 3, pp. 213-224

[21] WEBSTER, R. & OLIVER, M.A., 1990: Statistical methods in soil and land survey. -Oxford University Press, Oxford, 316 pp.

[22] WMO (Water Company of Overijssel), 1981: Grondwaterwinning diepe pakket in de omgeving van Schalkhaar (inventarisatieonderzoek)